Using SVMs and discriminative models for speech recognition

Nathan Smith and Mark Gales

May 2002

Cambridge University Engineering Department

ICASSP 2002

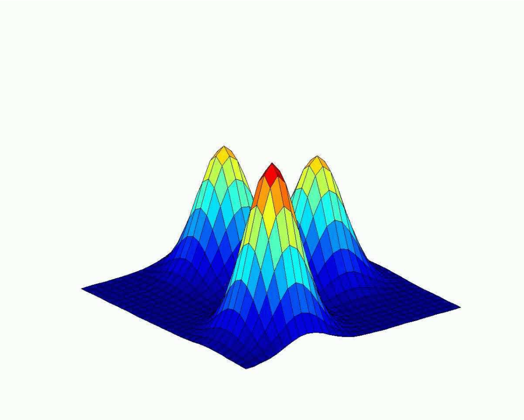
Overview

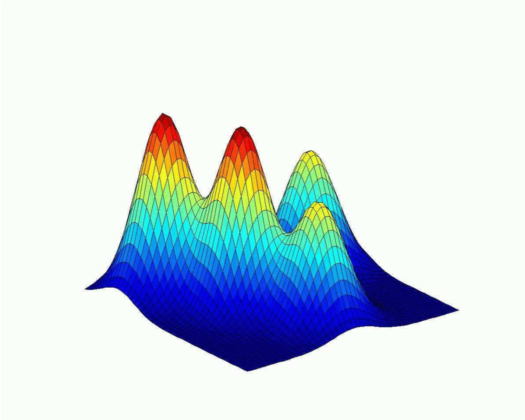
- MAP decoders
- Practical decoders
- Improving practical decoders - score-spaces
- Experiments and Results
- Conclusions

MAP decoder

utterance $\boldsymbol{O} = (\boldsymbol{o}_1, \ldots \boldsymbol{o}_t, \ldots \boldsymbol{o}_T), \boldsymbol{o}_t \in \mathbb{R}^D$

acoustic model class ω_1





acoustic model class ω_2

$p(\boldsymbol{O}|\omega_1)$

assume equal class priors

$p(\boldsymbol{O}|\omega_2)$

$$\rho(\boldsymbol{O}) = \ln p(\boldsymbol{O}|\omega_1) - \ln p(\boldsymbol{O}|\omega_2)$$

choose ω_1 if $\rho(\boldsymbol{O}) > 0$, else ω_2

MAP decision rule

Only optimal if models are correct

Practical MAP decoders

Correct acoustic models rarely known - model estimates may be

- Generative

Model data density Trained generatively e.g. MLE Implicitly train decision boundaries

- Discriminative

No assumptions of model correctness Trained discriminatively e.g. MMIE Explicitly train decision boundaries

- Model estimates give suboptimal performance ...
- ... but can we extract extra useful information from them?

Score-spaces - theory

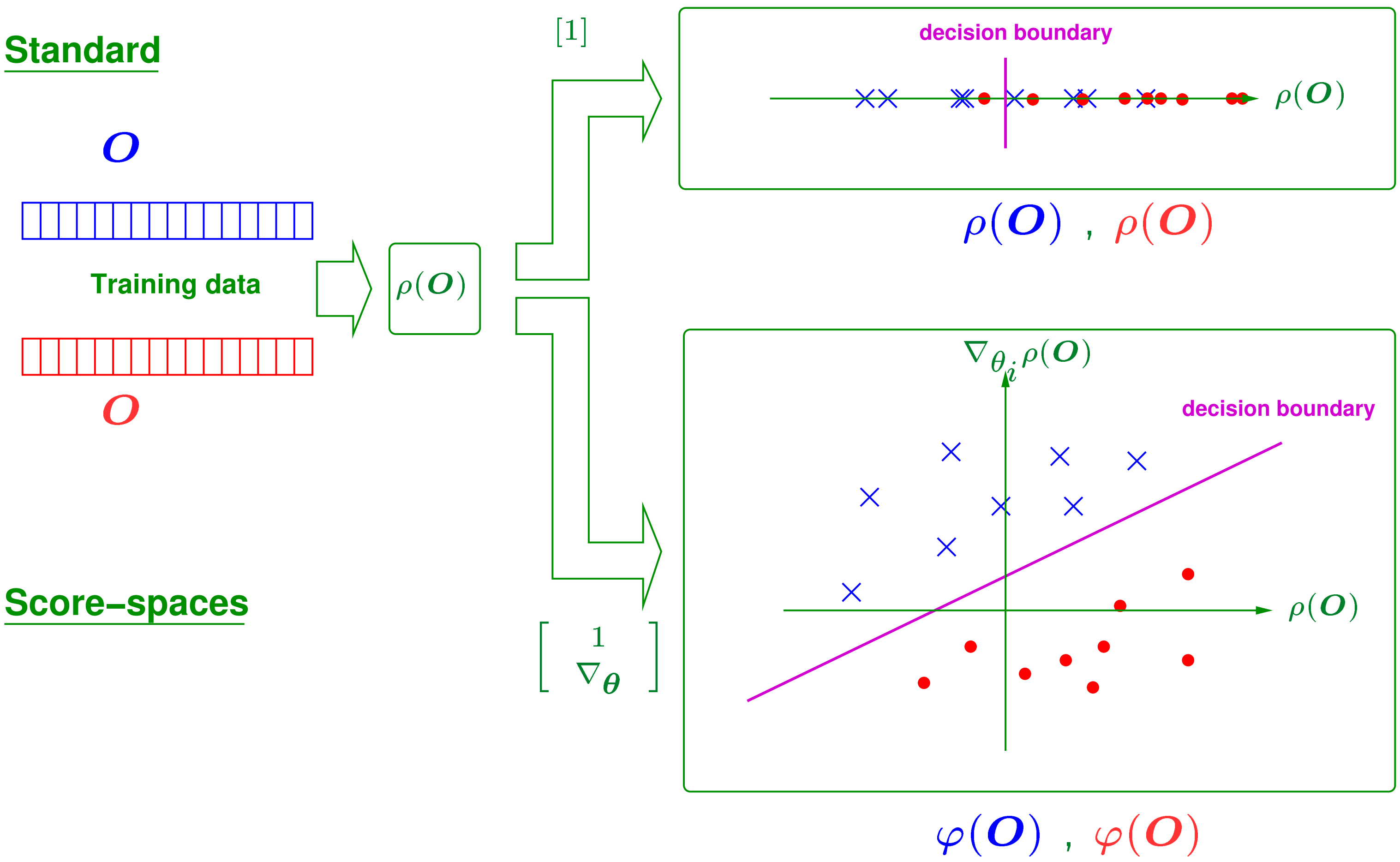
- Correct and estimated models have same functional form
- Correct models: log likelihood-ratio $\rho(\boldsymbol{O})$, parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top]^\top$
- Estimated models: log likelihood-ratio $\hat{\rho}(\boldsymbol{O})$, parameters $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\theta}}_2^\top]^\top$

Expand $\rho(\boldsymbol{O})$ about $\hat{\boldsymbol{\theta}}$ using Taylor series,

$$\begin{aligned}
\rho(\boldsymbol{O}) &= \begin{bmatrix} 1 \\ (\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1) \\ (\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2) \end{bmatrix}^\top \begin{bmatrix} \hat{\rho}(\boldsymbol{O}) \\ \nabla_{\hat{\boldsymbol{\theta}}_1}\hat{\rho}(\boldsymbol{O}) \\ \nabla_{\hat{\boldsymbol{\theta}}_2}\hat{\rho}(\boldsymbol{O}) \end{bmatrix} + O((\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2) \\
&= \boldsymbol{w}^\top \varphi(\boldsymbol{O}) + O((\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2)
\end{aligned}$$

- Optimal MAP decoder is $\rho(\boldsymbol{O}) = 0$
- Approximated by linear classifier in $\varphi(\boldsymbol{O})$
- So map each $\boldsymbol{O}$ to $\varphi(\boldsymbol{O})$, then estimate $\boldsymbol{w}$
- $\varphi(\boldsymbol{O})$ is a score-space

Discrimination in score-space



Classifiers in score-space

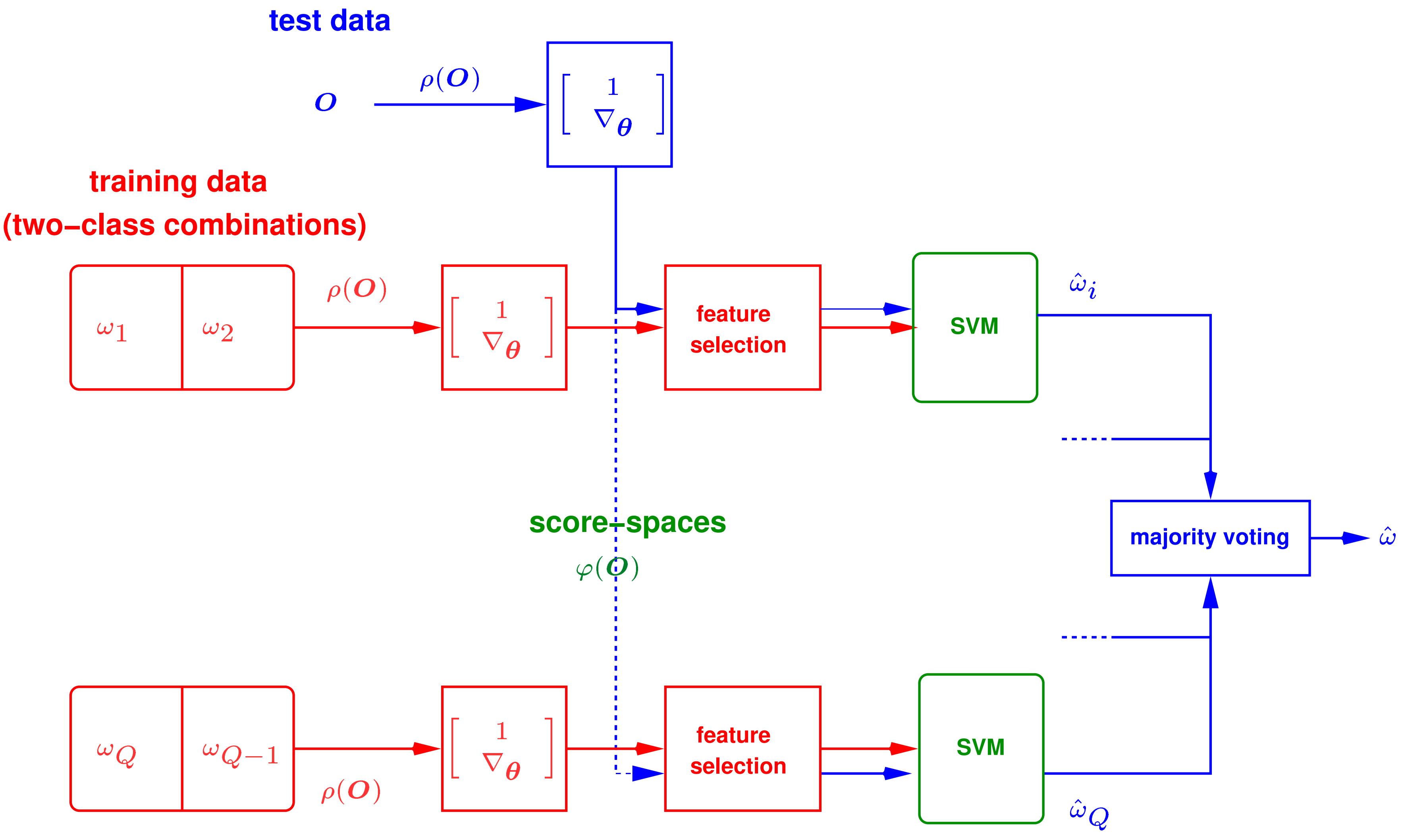
Use SVMs because they . . .

- prevent over-training in high-dimensional spaces
- produce linear/non-linear decision boundaries
- provide best chance of recovering optimal MAP decoders?

But SVMs . . .

- are binary classifiers

Multiclass SVM classifier in score-space



Experimental Setup

Task:

- Speaker-independent isolated letter database (ISOLET)
- E-set letters only {B C D E G P T V Z}
- 240/60 training/test utterances per letter

Acoustic models trained:

- CDHMM, GMM state output distributions
- 10-emitting states for letters, 1-emitting state for silence
- Feature-vector = (12 MFCCs + log energy + delta + acceleration)
- Three training regimes: ML, MMI-5, MMI-20

Classifiers tested:
'Standard' minimum Bayes error scheme SVMs in score-spaces defined on derivatives of means

Results

Standard schemes: lowest test error rates are ...

models	num comp per state	test error rate
ML	4	6.7
MMI-5	4	5.0
MMI-20	6	4.3

SVMs in score-space: models with 2 comp per state ...

models	test error rate		
	classifier		
	standard	SVM	(dim)
ML	8.7	3.7	(200)
MMI-5	5.4	3.2	(350)
MMI-20	4.8	4.1	(400-600)

... statistical significance?

Conclusions and further work

Preliminary results suggest score-spaces

- extract extra discriminatory information from acoustic models
- give good approximations to optimal MAP decoders
- give classifiers which compare well with MMI classifiers
- (full alphabet, ML models: 2.1% error rate)

Still to come ...

- larger tasks
- continuous speech recognition?

Any suggestions ...

Thanks to Phil Woodland, Dan Povey and KK Chin (Cambridge University)